Data Wrangling/ EDA Project Submission 1

Caroline Sullivan, Joe Thompson, Genevieve Purcell, Sasha Porter

1. **What is in your data?**

   Our dataset contains voting information for Virginia's presidential elections from 2000 to 2020. It includes the election year, the state (which is exclusively Virginia), and detailed data for each county, including their FIPS code (a county-specific identifier), the number of votes cast for each presidential candidate, and the total votes in each county. The candidates are categorized by political party—Democrat, Republican, and Other—providing a clear breakdown of the political landscape during each election.

2. **How will these data be useful for studying the phenomenon you're interested in?**

   Historical voting data is often a reliable predictor of future elections, and this dataset will be crucial for forecasting the 2024 presidential outcome in Virginia, both at the county level and for the state as a whole. By providing detailed information on the distribution of votes among candidates and parties from previous elections, the data allows for a granular analysis. With the results broken down by county, we can examine voting patterns across regions, identifying trends like consistent party leanings or significant shifts in voter behavior across election cycles. Leveraging these patterns from prior years will enable us to make more accurate predictions for 2024. Additionally, by analyzing total vote counts over time, we can gain valuable insights into voter turnout trends, helping to refine our predictions even further.

3. **What are the challenges you've resolved or expect to face in using them?**

   In 2020, we observed a significant change in how the county's presidential votes were recorded. Unlike previous years, where votes were categorized as "TOTAL" votes, the data was segmented into absentee, election day, and provisional votes, rather than presented as a single total. We believe this shift was likely influenced by COVID-19, which limited in-person voting options. To ensure consistency across all years, we will likely need to use Python to aggregate these different vote types for 2020, so that each year's data aligns with a uniform format.

   A challenge we encountered is that for the year 2000, the Green Party is listed separately, alongside the Democratic, Republican, and Other categories. Similarly, for 2020, the Libertarian Party appears as a distinct category. To maintain consistency in formatting across all years, we recommend consolidating both the Green and Libertarian parties under the "Other" category. This will ensure uniformity in party classification throughout the dataset.